Creation of a Topic Maps-Based Wiki with an Article Similarity Measurement

Shu Matsuura[a*], Motomu Naito[b] & Hiromi Toyota[c]
[a]*4-1-1 Nukuikita, Koganei, Tokyo 184-8501, Tokyo Gakugei University, Japan*
[b]*3-3-1 Yshiki-cho, Takahama, Aichi 444-1331, Knowledge Synergy Inc., Japan*
[c]*3350 Honmachida, Machida, Tokyo 194-0032, Honmachida Higashi Elementary School, Japan*
*shumats0@gmail.com

Abstract: We created a pilot case of a Topic Maps-based wiki site to exchange ideas about children's behavior in elementary schools. The topic map of this site consisted of the article and subject topic types. The subject topic type consisted of topics classified into areas such as "behavior," "competence," "field," and "school time." Each article was registered as an instance of the article topic type and associated with relevant subject topics. To measure the similarity between two articles and to relate articles on the basis of the similarity, we used the Tanimoto Similarity. To improve similarity-based retrieval, it was suggested that more specific subject topics characterize the articles.

Keywords: Topic Maps, Wiki, Tanimoto Similarity

Introduction

Since 1998, it has been indicated that elementary schools in Japan often face difficulties in managing classes. [1]. To overcome this situation, both experienced and inexperienced teachers need to share the various problems they face in the classrooms and their solutions. For this purpose, we created a pilot case of the wiki site [2] based on Topic Maps (ISO/IEC JTC1/SC34) [3]. The topic represents the subject. The topics are interconnected through their associations. Information resources are linked to the corresponding topic by the occurrences [4]. Topic Maps fits to the construction of a wiki site [5, 6].

Our wiki site aims to let teachers who write or read articles find information related to it. In general, while an article represents at least one subject, it can be related to various contexts, even if the writer has little intention of doing this. In other words, if two articles share certain common factors, they are similar. Reading articles that have certain degree of similarity may evoke some hints on solving the problem from a different aspect.

This paper introduces a simple method for associating articles. We classify articles as individual topic instances. Then, we ask writers to choose topic names related to the article. Thus, individual articles are characterized by sets of chosen topic names. We evaluate the similarity between any two articles by calculating Tanimoto Similarity.

1. Method

1.1 Educational issues in children's daily life

In this section, we list educational issues faced by elementary school children. We use these particular issues as topic instances to associate with wiki articles.

Competence. Children's education aims to have children attain social skills. OECD has defined the key competency for living and cooperating in the modern society from a broader standpoint [7]. We consider five categories of children's competency—physical strength, intelligence, willingness, practice, and communication. Table 1 shows the indications for assessing the competence of the five categories. These indications have been selected through discussions of elementary school teachers led by one of the authors.

Table 1. Children's competency in school life.

Category	Competency types	Indications
Physical strength	Posture	*Standing and sitting postures*
	Group gymnastic skill	*Ball games, group games*
	Individual gymnastic skill	*Run, jump, apparatus gymnastics*
	Health	*Likes and dislikes in food, illness*
Intelligence	Reading skill	*Reading aloud, comprehension, Kanji*
	Writing skill	*Writing letters, figure, sentences*
	Arithmetic skill	*Arithmetic*
	Logical thinking	*Vocabulary, comprehension, thinking, expressing*
Willingness	Expressiveness	*Smiling, laughing*
	Perseverance	*Tenacity*
	Will for living	*Positive thinking, not depressed*
	Will to act	*Positive attitude*
Practice	Cooperation with friends	*Play with friends*
	Cooperation in work	*Act in cooperation*
	Roles in daily life	*Day duty, activity in charge*
	Rules in daily life	*Following rules*
Communication	Listening to	*Looking at, listening to others*
	Assessment of situation	*Behave according to the situation*
	Expression	*Tell others what the child feels or thinks*
	Sympathy	*Guess what others feel*

Behavior. Children's problems appear in their behaviors. Indications in the behavioral types consist of *school refusal*, *truancy*, *antisocial behaviors*, *rude*, *nonsense*, *forgotten*, *perverseness*. They supplement the assessment of competency. Describing children's behavior is required particularly when the problem is not directly attributable to any specific competency factor.

Space and Time. Many of the children's problematic behaviors occur at a specific time and place. We categorized the typical time periods of children's activities in Japanese schools into the following 8 types, as "going to school," "morning assembly," "classrooms," "intermissions," "school lunch," "cleaning rooms," "social activities at school," and "getting out of school". Under these time period types, 26 indications of children's activities were located in total.

The category "field" consists of 7 location types and 14 indications. The types of "home," "school," and "classroom" specify physical location, while the types "with friends," "with relatives," "local community," and "external community" indicate that the fields are characterized by human relationships. The classified items are not mutually exclusive, but instead supplemental in characterizing the children's environment.

1.2 Constructing the topic map

We constructed an RDBMS topic map and a website on the basis of a topic map development suite Ontopia 5.1.0. [8]. Our topic map consists of "article" and "subject" topic types. The article type contains individual articles written by wiki users.

The article type is associated with the following four types by the broader_narrower association. The “cares_article” type includes instances of articles on teacher’s troubles or worries. The “rules_article” type includes useful rules that are applied in the class and work to solve these problems. The “suggestions_article” type includes hints, ideas, and experiences that solve problems faced in class management. Finally, the “teachers_word_article” type includes teachers’ words that either were beneficial or, in contrast, not useful in solving the children’s problem.

When an article of one of the four article types was uploaded, an article instance of the type is created. Posted article automatically gets a unique base name and a public subject identifier. Finally, the contributor associates his or her article with the subject topics by the “article_related_with_subject” association, to characterize the meaning of the article.

For subject topic type, we subdivided the type into two categories of the “article_subject” and the “article_situation”. The article_subject was further divided into four types of “competence,” “behavior,” “school time,” and “field,” which were described in Sec. 1.1. The subject topic instances correspond with indications described in the above section.

The “article_situation” is further divided into “teacher’s_reflection” and “worry.” The “teacher’s_reflection” type concerns with the instructor’s attitude toward a child or a class, including 5 topic instances of “child_assistance_mind,” “class_assistance_mind,” “class_management,” “educate_child,” and “watching_child.” The “worry” topic concerns who is particularly worrying, including 3 instances of “children_worry,” “parents_worry,” and “teacher_worry.”

1.3 Similarity calculation

In this system, wiki writers characterize their articles by the set of subjects associated with the article. Although this is an indirect method of characterizing articles, it offers a simple approach to measure the similarity between two articles. We regard the two articles as having similar features if they have common subjects associated with them.

To measure the similarity, we calculate the Tanimoto Similarity between the sets of subjects associated with the articles. Tanimoto Similarity is the rate of intersection of the union of two sets, which assigns “1” for equivalence and “0” for no similarity [9]. If we assign the set of associated subjects A of an article a and the set of subjects B of an article b, the Tanimoto Similarity T_{ab} between articles a and b is written as $T_{ab} = |A \cap B| / |A \cup B|$.

2. Results and Discussion

At present, a total of 102 articles have been written: 67 articles for the cares topic, 14 for the suggestion topic, 14 for the rules topic, and 7 for the teachers_words topic.

Figure 1a shows a semi-log plot of the histogram showing the similarity between any two articles; 64.7% of the articles had fewer than 9 associations. The frequency of the article pairs decayed almost logarithmically as their similarity increased. The articles having 3 or 4 associations were the most frequent, although many associations were mutually exclusive and could not be selected for the same article. 33.5% of the pairs showed a similarity ranging from 0.1 to 0.3. In this study, the range of similarity for showing similar articles is set to more than or equal to 0.3. This value covers only 5.6% of the article pairs.

Here we consider a simplified uniform model, in comparison with our wiki. We consider the articles $A = \{A_1, \ldots, A_N\}$, where A_i has m_i ($\leqq M$) associated subjects. Here M is the total number of associated subjects, and the number m_i of associated subjects is chosen uniformly at random from 0 to M. The subjects are chosen randomly from the set of subjects

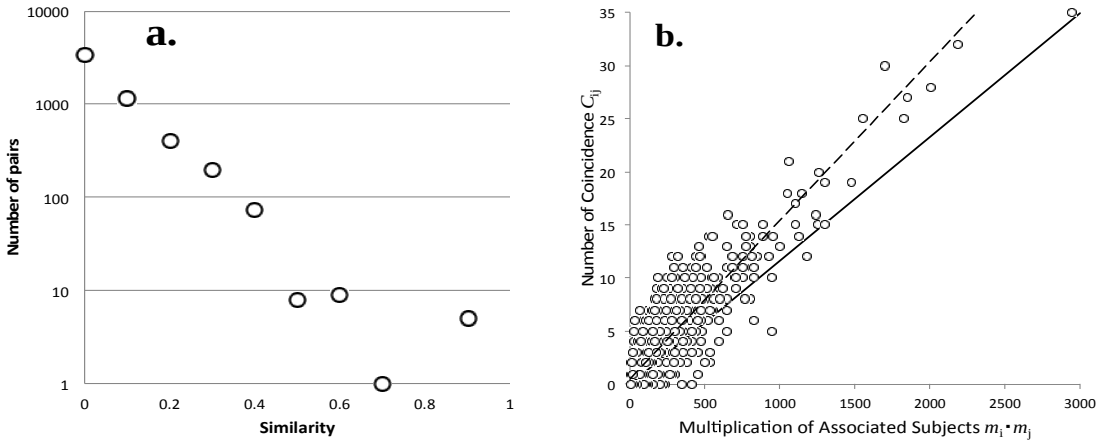

Fig. 1a, b. Left, a, semi-log plot of histogram of the similarity of article pairs. Right, b, plot of C_{ij} verses ($m_i \cdot m_j$) of pairs of actual wiki articles. A bold solid line indicates the slope of M^{-1} = 1/86. The dashed line indicates a fitting line whose slope is approximately 1/67.

S, where $S = \{s_1, \ldots, s_M\}$, for every A_i. Similarity between A_i and A_j is calculated by applying the Tanimoto Similarity on the sets of associated subjects m_i and m_j.

We define the number of equivalent subjects in the sets of associated subjects as "the number of coincidence C_{ij}." The number of coincidence C_{ij} between A_i and A_j is expected proportional to the multiplication $m_i \cdot m_j$, where the coefficient of proportion will be M^{-1}. This relationship is expressed by an equation $C_{ij} = M^{-1}(m_i \cdot m_j)$ (1).

Figure 1b shows a plot of C_{ij} verses $m_i \cdot m_j$ on the actual wiki articles. The solid line shows eq. (1), where M is the real number of subject instances; M = 86. The plotted points are concentrated at lower values of $m_i \cdot m_j$. In addition, the plotted points show higher values of C_{ij} as compared with the line of eq. (1) with M = 86. The dashed line shows the relationship with M = 67. This implies that the range of selection of associated subjects was effectively smaller than the possible number of choices. To increase the number of articles having high similarity, we have to advise wiki authors to associate a single subject with as many viewpoints as feasible.

Finally, we consider the relevance of subject topics in specifying the articles. Table 5 shows six subtypes of subject topics and their actual usage in characterizing the articles. While the Article_situation types have only a few instances, larger numbers of articles are associated per instance, in comparison with the Article_subject types.

Table 2. Number of subject topic instances associated with articles.

Subject type	Subject sub-type	Number of instances selected by contributors	Mean number of associated articles
Article _subject	Behavior	7	6.4
	Competence	35	8.1
	School_time	26	8.5
	Field	8	14.8
Article _situation	Teacher's_reflection	5	19.4
	Worry	3	31.0

In general, an article consists of several specified conditions such as "who is it about," "what," "when," "where," "why," and "how did it happen." These factors are associated with the instances of Article_subject type. Thus, in many cases, the Article_subject instances are regarded as narrower than the article topic instances.

On the other hand, the instances of Article_situation subject type do not characterize what is specifically described in the article. Rather, these instances classify the articles from a broader perspective. Then, the instances of Article_situation subject type are associated with many article topic instances.

To measure the similarity in the articles' specificity, the similarity measurement for the narrower subjects is more effective than that for the broader ones. Improvement in the interface is required to increase assignment of narrower subjects and at the same time decrease the burden of checking many items.

Our approach in this study was to let contributors reconsider what subjects their text could be associated with even though such relationships was not explicitly written in their texts. Another important analysis method is obviously the text mining, which explore new trends from large number of texts [10, 11]. The text mining methods will be effective to extract trends in the wiki articles if the sufficient number and variety of articles are collected.

3. Conclusion

We created a pilot case for constructing a wiki site based on Topic Maps to share the problems, suggestions, and effective rules concerning the life of elementary school children. The articles were stored as article topic instances and were associated with various subject topics including children's behaviors, core competencies, and life skills. On the basis of the retrieval of associated topics, the articles were automatically organized and rendered searchable in the wiki site. In addition, the Tanimoto Similarity calculation evaluated the similarity of subjects associated between two articles. Because the contributors' choice of associated subject influences the effect of this method, the interface needs to be improved for appropriate choice of subject with low burden.

Acknowledgements

This study has been funded by a Grant-in-Aid for Scientific Research (C) 21500842 from the Ministry of Education, Culture, Sports, Science and Technology, Japan.

References

[1] National Institute for Educational Policy Research. http://www.nier.go.jp/shido/centerhp/unei.pdf (in Japanese).
[2] Wiki site of this work, http://tm.u-gakugei.ac.jp/ca/k/ (in Japanese).
[3] JTC 1/SC 34. http://www.itscj.ipsj.or.jp/sc34/.
[4] Pepper, S. The TAO of Topic Maps. http://www.ontopia.net/topicmaps/materials/tao.html#d0e632.
[5] Cerny, R. (2008). Topincs Wiki – A Topic Maps Powered Wiki. *Lecture Notes in Computer Science*, 4999/2008, 57-65.
[6] Thomas, H. (2007). Design Principle for a Topic Maps Wiki –The Wiki Way of Topic Maps. http://www.informatik.uni-leipzig.de/~tmra/2007/slides/redmann_TMRA2007.pdf.
[7] OECD. The Definition and Selection of Key Competencies, Executive Summary. http://www.oecd.org/dataoecd/47/61/35070367.pdf.
[8] Ontopia code google. http://code.google.com/p/ontopia/.
[9] Rogers, D. J. & Tanimoto, T. T. (1960). A Computer Program for Classifying Plants. *Science* 132, 1115-1118.
[10] Hearst, A. M. (1999). Untangling Text Data Mining, *Proc. of ACL-99*, 3-10.
[11] Feldman, R. & Sanger, J. (2006). *The Text Mining Handbook*, Cambridge University Press.